

# Distribution of Mutual Information

Henry D. I. Abarbanel<sup>†</sup>

*Department of Physics and Marine Physical Laboratory, Scripps Institution of Oceanography,  
University of California, San Diego, La Jolla, CA 92093-0402*

Naoki Masuda<sup>‡</sup>

*Department of Physics and Institute for Nonlinear Science,  
University of California, San Diego, La Jolla, CA 92093-0402*

M. I. Rabinovich<sup>§</sup>

*Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA 92093-0402*

Evren Tumer<sup>\*\*</sup>

*Department of Physics and Institute for Nonlinear Science,  
University of California, San Diego, La Jolla, CA 92093-0402*

(Dated: February 8, 2008)

In the analysis of time series from nonlinear sources, mutual information (MI) is used as a nonlinear statistical criterion for the selection of an appropriate time delay in time delay reconstruction of the state space. MI is a statistic over the sets of sequences associated with the dynamical source, and we examine here the distribution of MI, thus going beyond the familiar analysis of its average alone. We give for the first time the distribution of MI for a standard, classical communications channel with Gaussian, additive white noise. For time series analysis of a dynamical system, we show how to determine the distribution of MI and discuss the implications for the use of average mutual information (AMI) in selecting time delays in phase space reconstruction.

PACS numbers: PACS Numbers: 84.40.Ua, 89.70.+c, 05.45.-a, 05.10.L, 05.45.Tp

Information theory [1, 2, 3] characterizes general dynamical systems through a nonlinear connection between sequences of symbols associated with action of the system. The connection could be between inputs and outputs of a channel or could be associated with predictions of future measurements from past observations [4]. Shannon's identification of MI as the essential statistic in such systems gives a framework for the discussion of applications as diverse as fiber optic communications systems with time scales shorter than nanoseconds and nervous systems with time scales of a few tens of milliseconds to seconds.

In the analysis of time series coming from nonlinear sources [5, 6] one reconstructs a proxy state space using observations of a single variable  $V(t)$ . A useful reconstructed state space employs the observed variable  $V(t)$  and its time delays to form a data vector

$$\mathbf{y}(t) = [V(t), V(t - T\tau_s), \dots, V(t - (d-1)T\tau_s)], \quad (1)$$

where  $\tau_s$  is the sampling time and  $T$  is an integer. We must choose  $T$  so that the components of  $\mathbf{y}(t)$  are 'inde-

pendent enough' of each other to be good coordinates for the space. For this purpose

$$I(T) = \sum_{\{V(t), V(t+T\tau_s)\}} P(V(t), V(t+T\tau_s)) \log_2 \left\{ \frac{P(V(t), V(t+T\tau_s))}{P(V(t))P(V(t+T\tau_s))} \right\}, \quad (2)$$

the AMI, has been used. Following the suggestion of Fraser and Swinney [7],  $T$  is chosen where  $I(T)$  has its first minimum.

The AMI  $I(T)$  answers the question: how much, in bits, do we know about the measurement at  $t+T\tau_s$  from the measurement at  $t$ , averaged over the whole time series or attractor. However, it does not tell us how MI is distributed over the time series. We have only the hope that the distribution of MI is sharp enough to provide a choice for  $T$  which is useful over the whole time series.

If the dynamics of the process producing  $V(t)$  has two time scales, for example, one might expect the distribution of MI to reflect that by exhibiting two distinct peaks associated with each process efficiently connecting to it-

self through the dynamics of the system. In such a case the AMI may tell us little about how to select a value of  $T$  to use in making  $\mathbf{y}(t)$ . Since the MI whose mean is evaluated in Eq.(2) tells us how well we can predict  $V(t)$  knowing  $V(t-T)$ , the presence of two time scales leading to a multivalued distribution of MI is quite natural.

In such circumstances, it may not be a good idea to use the  $T$  chosen by the AMI, but to select different  $T$  in different parts of the attractor, or to use another prescription altogether for choosing  $T$ . As many physically or biologically interesting dynamical systems have two or more relevant time scales, the distribution of MI presents an interesting issue.

Time delay phase space reconstruction is widely used as an initial step in the analysis of time series from nonlinear sources, so it is useful to establish how MI, considered as a 'statistic' distributed over the measurements, is distributed. This will allow one to proceed well beyond the use of the average alone.

Our considerations here may provide a framework for efficient use of a broad set of communications channels. Presented with a channel, which need not be stationary, one may wish to know those sequences which maximize MI to allow the best use of that channel. Those sequences need not always correspond to the AMI associated with the channel. In other words the distribution of MI can depend both on the channel and on the symbol sequences conveyed by the channel.

MI provides a nonlinear relation between one sequence of symbols and second sequence of symbols. These could be the input and output, respectively, of a communications system or of a network of neurons performing a functional task. We consider a sequence  $S = \{..., s(-1), s(0), s(1), ...\}$  observed at discrete times and ask about its nonlinear connection to a second sequence  $R = \{..., r(-1), r(0), r(1), ...\}$ . The symbols  $\{r(l)\}$  and  $\{s(k)\}$  may take discrete or continuous values.

We are interested determining the properties of  $S$  from measurements of the sequence  $R$ . The ability to do this is characterized by the MI

$$I(s(k), r(l)) = \log \left\{ \frac{P_{SR}(s(k), r(l))}{P_S(s(k))P_R(r(l))} \right\}. \quad (3)$$

$P_{SR}(s, r)$  is the joint distribution function of symbols  $s$  taken from the input sequence  $S$  and symbols  $r$  taken from the output sequence  $R$ .  $P_{SR}(s, r)$  is the essential ingredient in determining MI. It depends both on the sequences  $S$  and  $R$  and, importantly, on the channel or dynamical process connecting the  $s(k)$  to the  $r(l)$ .  $P_S(s) = \sum_r P_{SR}(s, r)$  and  $P_R(r) = \sum_s P_{SR}(s, r)$  are the distributions of  $S$  symbols and  $R$  symbols, respectively. When the sequences are independent,  $P_{SR}(s, r) = P_S(s)P_R(r)$ , and  $I(s, r) = 0$ . The MI  $I(s, r)$  is a variable over  $(s, r)$ , and it has its own distribution function  $P_I(x)$ .  $P_I(x)$  tells us the frequency with which the value

$x = I(s, r)$  occurs.

The distribution of MI is defined by

$$P_I(x) = \int ds dr P_{SR}(s, r) \delta(x - I(s, r)), \quad (4)$$

with  $I(s, r) = \log \left\{ \frac{P_{SR}(s, r)}{P_S(s)P_R(r)} \right\}$ .  $\langle x \rangle = \int dx x P_I(x) = \int ds dr P_{SR}(s, r) I(s, r)$ , as it should be. While  $\langle x \rangle > 0$ , negative values of  $I(s, r)$  are possible [2]. Negative values of  $x = I(s, r)$  reflect the circumstance that  $(s, r)$  pairs may occur less frequently than the individual symbols themselves. If a process produces out-of-phase appearances of  $r$  and  $s$ , for example, negative values of MI will occur.

To evaluate  $P_I(x)$  using the observed  $P_{SR}(s, r)$  we could solve the delta function condition  $x = I(s, r)$  for, say,  $s_*(x, r)$  and then perform the integral over  $r$ . An alternative method is to Fourier transform  $P_I(x)$  giving  $Q_I(f) = \int dx e^{-2\pi i f x} P_I(x) = \int ds dr e^{-2\pi i f I(s, r)} P_{SR}(s, r)$ .  $P_I(x)$  is recovered by inverse Fourier transform  $P_I(x) = \int df e^{i2\pi f x} Q_I(f)$ . In effect we are using the Fourier variable  $f$  as a Lagrange multiplier to implement the required delta function on  $x = I(s, r)$ . We avoid solving for  $s_*(x, r)$  and require only an integral over the same ingredients used in evaluating the AMI.

A classical example is provided by the Gaussian distribution in  $(r, s)$

$$P_{SR}(s, r) = \frac{\sqrt{ab - \sigma^2}}{\pi} e^{-(ar^2 + bs^2 + 2\sigma sr)}. \quad (5)$$

The quantity  $\xi = \frac{\sigma^2}{ab} < 1$  for this to be normalizable.  $P_R(r) = \sqrt{\frac{ab - \sigma^2}{\pi b}} e^{-r^2(1 - \xi)}$ ,  $P_S(s) = \sqrt{\frac{ab - \sigma^2}{\pi a}} e^{-s^2(1 - \xi)}$ , and  $I(s, r) = X_0 - (\frac{\sigma^2 s^2}{a} + \frac{\sigma^2 r^2}{b} + 2\sigma sr)$ , where  $X_0 = -\frac{1}{2} \log(1 - \xi)$ . When  $\sigma = 0$ ,  $I(s, r) = 0$ .

From this we find  $Q_I(f)$

$$Q_I(f) = e^{-i2\pi f X_0} \frac{1}{\sqrt{1 + 4\pi^2 f^2 \xi^2}}. \quad (6)$$

As expected, when  $\xi = 0$ ,  $Q_I(f) = 1$  and  $P_I(x) = \delta(x)$ .  $P_I(x)$  can also be evaluated for  $\xi \neq 0$

$$P_I(x) = \frac{1}{\pi \xi} K_0 \left( \frac{X_0 - x}{\xi} \right), \quad (7)$$

where  $K_0(z)$  is a zeroth order modified Bessel function [8].  $K_0(z)$  is symmetric in  $z$  and has a logarithmic singularity at  $z = 0$ . The AMI is  $X_0$ ,  $\langle (x - X_0)^2 \rangle = \xi^2$ , and  $\langle (x - X_0)^{2m} \rangle = \xi^{2m} ((2m - 1)!!)^2$ .

A standard example of a Gaussian channel is given by the case of a Gaussian input signal with  $P_S(s) = \frac{1}{\sqrt{2\pi}S} e^{-\frac{s^2}{2S^2}}$ , and with conditional probability of response  $P_{SR}(r|s) = \frac{1}{\sqrt{2\pi}N} e^{-\frac{(r-s)^2}{2N}}$ . This represents a statistical

signal transported through a passive channel with additive, Gaussian, white noise [2]. The signal to noise ratio in this channel is  $\frac{S}{N}$ .

The distribution of mutual information is that just given with  $a = \frac{1}{2N}$ ;  $b = \frac{1}{2N}(1 + \frac{N}{S})$ ;  $\sigma = \frac{1}{2N}$ , leading to the AMI  $X_0 = \frac{1}{2} \log(1 + \frac{S}{N})$ , which is familiar [2], and moments, which are a new result, given as above with  $\xi = \frac{\frac{S}{N}}{1 + \frac{S}{N}}$ . This means that for a large  $S/N$ , the AMI grows logarithmically in  $S/N$ , but, perhaps surprisingly, the moments about this average are order unity or larger. That is the distribution is not sharp, but becomes broad. For  $\frac{S}{N}$  small,  $X_0 \approx \frac{S}{2N}$  and the moments are powers of  $\frac{S}{N}$ . In fact, for small  $\frac{S}{N}$  or noisy channels, the RMS value of MI is twice the mean value.

Other  $P_{RS}(r, s)$  must be dealt with numerically, and the sequence values must be discrete. From  $P_{SR}(s(k), r(l))$  a MI matrix is generated:

$$I(s(k), r(l)) = \log_2 \left( \frac{P_{SR}(s(k), r(l))}{P_S(s(k)) P_R(r(l))} \right), \quad (8)$$

and  $Q(f)$  is approximated by

$$\begin{aligned} Q(f) &\approx \sum_{s(k), r(l)} P_{SR}(s(k), r(l)) e^{-i2\pi f I(s(k), r(l))} \Delta r \Delta s \\ &= \sum_k P_I(x_k) e^{-2\pi i f x_k} \Delta x_k, \end{aligned} \quad (9)$$

where  $\Delta s$  and  $\Delta r$  are the size of the symbol bins.  $P_I(x)$  is evaluated bins of size  $\Delta I$ . We denote  $p_m = P_I(x_m) = P_I(m\Delta I + I_{min})$ , where  $m$  is an integer that ranges from 0 to  $N_{bin}$  and  $I_{min}$  is the minimum value of MI.  $p_m$ , [9] is determined for  $N_{bin}$  values of  $f_n = \frac{n}{N_{bin}\Delta I}$ .  $n$  is an integer in the range  $[-\frac{1}{2}N_{bin}, \frac{1}{2}N_{bin}]$ . The approximation to  $Q(f_n)$  is then  $Q(f_n) = \Delta I e^{-2\pi i I_{min} f_n} \sum_{k=0}^{N_{bin}} p_k e^{-2\pi i k \Delta I}$ , so the inverse Fourier transform of  $\frac{Q(f_n)}{\Delta I} \exp\left(\frac{2\pi i I_{min} n}{N_{bin}\Delta I}\right)$  is  $p_m$ .

To apply this to the MI used in selecting the time delay in phase space reconstruction we identify  $s(t) = V(t_0 + k\tau_s) = s(k)$  and  $r(t) = V(t_0 + (l + T)\tau_s) = r(l)$  where  $t_0$  is some initial time in the data. We have investigated  $P_I(x)$  for various  $T$  for two simple dynamical systems: (1) the Lorenz system, and (2) a nonlinear hysteretic circuit using data provided by L. Pecora and T. Carroll (Personal Communication). Both data sets are discussed in [5].

For the Lorenz system data set we used  $2 \times 10^6$  values of the variable  $x(t)$ . The first minimum of AMI for these data is at  $T = 10$  in units defined by the numerical integration step. We evaluated  $P_I(x)$  for various  $T$ . In Figure 1 we show this distribution for  $T = 1$ , for  $9 \leq T \leq 11$ , and for  $T = 16$ . We can see that for  $T = 1$   $P_I(x)$  has a single well defined peak, indicating little variation of MI across the attractor. The AMI for this distribution is larger

than that for  $T \approx 10$ . This is the usual [7] indication that  $T = 1$  is not a preferred choice. For  $9 \leq T \leq 11$   $P_I(x)$  is nearly the same with two well formed, but nearby, peaks. These we attribute to fast and slow motions on the attractor associated with motion about the unstable fixed points and motion through the “neck” near the original in phase space. As the peaks are rather close, the use of a single  $T$  to provide coordinates for the data vector  $\mathbf{y}(t)$  over the whole attractor would appear to be a good construct. If one wished to resolve the two time scales exposed by the peaks in  $P_I(x)$ , the use of two different time delay coordinate systems in different parts of the attractor would be appropriate.

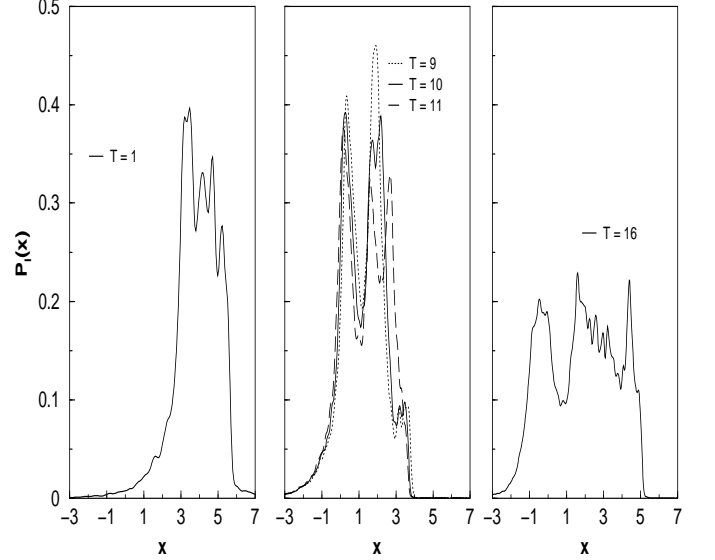


FIG. 1: Distribution of MI for the Lorenz system for various  $T$  used in the data vector  $\mathbf{y}(t)$ . The left panel has  $T = 1$ , the middle panel  $9 \leq T \leq 11$ , and the right panel has  $T = 16$ . The AMI for  $T = 1$  is larger than for  $T \approx 10$ . For  $T = 1$  and for  $9 \leq T \leq 11$  we have rather sharp distributions  $P_I(x)$ , so the use of a single value of  $T$  across the attractor is useful. For  $T \approx 10$  the bimodal nature of  $P_I(x)$  suggests that one could identify two processes contributing to  $P_I(x)$  and seek a value for  $T$  for each of them. For  $T = 16$  the distribution is so broad that no particular meaning can be associated with the AMI. This  $P_I(x)$  indicates many different processes contribute over the attractor, and thus, this is not a good value of  $T$  to use in forming the data vector.

For larger  $T$ , here  $T = 16$ , we observe substantial broadening in  $P_I(x)$ . Here the mutual information between  $V(t)$  and  $V(t + T\tau_s)$  is more or less uniformly distributed over the attractor suggesting all points are rather decorrelated, in an information theoretic sense, from all others. This means that all dynamical processes on the attractor convey little information from time  $t$  to time  $t + T$ . Equivalently, coordinates of the time delay data vector  $\mathbf{y}(t)$  formed with  $T = 16$  are so independent of each other as to constitute a bad choice for following

the dynamics of the system.

For the nonlinear circuit data we have 64,000 data points taken at  $\tau_s = 0.1$  ms. The minimum of AMI is  $T = 6$ . For these data  $P_I(x)$  is shown in Figure 2 for  $T = 1$ , for  $5 \leq T \leq 7$ , and for  $T = 11$ . We have a relatively sharp peak for  $T = 1$ , but high values of AMI. The peak narrows for  $T \approx 6$ , and then broadens again as  $T$  grows. We show  $T = 11$  where the broadening of  $P_I(x)$  seen already in the Lorenz system occurs again. This indicates that this value of  $T$  is not useful in reconstructing the entire attractor.

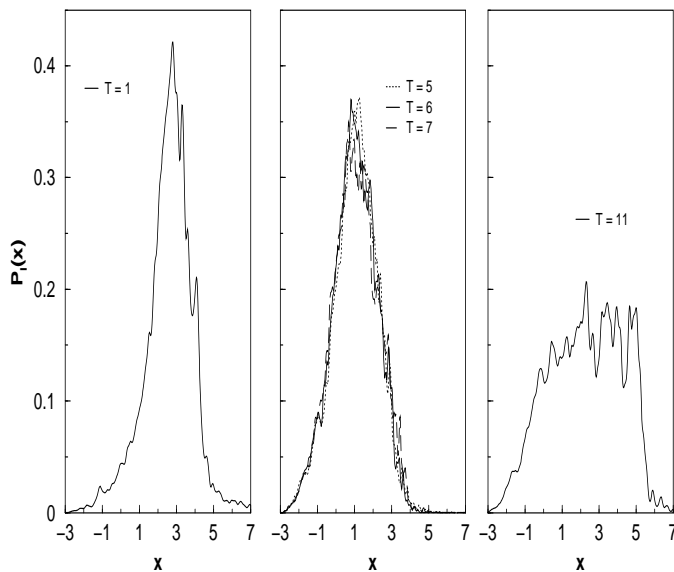


FIG. 2: Distribution of MI for the hysteretic circuit for various  $T$  used in the data vector  $\mathbf{y}(t)$ . The left panel has  $T = 1$ , the middle panel  $5 \leq T \leq 7$ , and the right panel has  $T = 11$ . The AMI for  $T = 1$  is larger than for  $T \approx 6$ . Each of these is a rather sharp distribution  $P_I(x)$ , so the use of a single value of  $T$  across the attractor is useful. For  $T = 11$   $P_I(x)$  has broadened indicating this value of  $T$  yields components of the data vector  $\mathbf{y}(t)$  which may be too independent for use.

$P_I(x)$  for various models and for a variety of experimental data will be reported in our larger paper [9]. In this short note we have introduced the distribution of MI and exhibited several examples, analytic and numerical, to illustrate its properties. The use of  $P_I(x)$  as illustrated here for understanding the distribution of MI in the connection between elements of the data vector  $\mathbf{y}(t)$  gives us a clearer understanding of the choice made some years ago by Swinney and Fraser [7] of using the first minimum of AMI to select  $T$ . It is not only low AMI among components of  $\mathbf{y}(t)$  that is important, but also one must have a distribution of MI which is sharp so that a single choice for  $T$  over the whole attractor can accurately capture the underlying dynamical processes.

We expect  $P_I(x)$  to have direct utility in the study of networks with active dynamical elements [9, 10] where the function of the network is to convey information from a set of source symbols  $s(k)$  to a set of response symbols  $r(l)$  with high fidelity. This fidelity is characterized by high values of MI between input and output, and maxima of  $P_I(x)$ , rather than the AMI, will indicate which processes  $s(k) \rightarrow r(l)$  are best communicated. These ideas will be fully explored in our larger paper [9].

This work was supported in part by the U.S. Department of Energy, Office Science, Division of Engineering and Geosciences, under grant DE-FG03-90ER14138, and in part by National Science Foundation grant NCR-9612250. Support was also received from the U. S. Army Research Office under contract No. DAAG55-98-1-0269; MURI Program in Chaotic Communications. ET acknowledges support from NSF Traineeship DGE 9987614.

<sup>†</sup> Institute for Nonlinear Science; Electronic address: hdia@jacobi.ucsd.edu

<sup>‡</sup> Electronic address: masuda@physics.ucsd.edu; Also Department of Mathematical Engineering and Information Physics, Graduate School of Engineering, University of Tokyo, Tokyo, 113-8656, Japan

<sup>§</sup> Electronic address: rabin@landau.ucsd.edu

<sup>\*\*</sup> Electronic address: evren@nye.ucsd.edu

- [1] C. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
- [2] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications* (MIT Press and John Wiley & Sons, 1961).
- [3] R. G. Gallager, *Information Theory and Reliable Communication* (John Wiley and Sons, New York, 1968).
- [4] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
- [5] H. D. I. Abarbanel, *The Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).
- [6] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, 1997).
- [7] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
- [8] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions* (Dover Publications, Inc., New York, 1965), chapter 5 by F. W. J. Olver. See formula, 9.6.21.
- [9] E. Tumer, N. Masuda, M. I. Rabinovich, and H. D. I. Abarbanel, *Characterizing information transport by the distribution of mutual information*, to be submitted to *Physical Review E*, Winter, 2000.
- [10] M. C. Eguia, M. I. Rabinovich, and H. D. I. Abarbanel, *Physical Review E* **62** (2000), to appear, November, 2000.